AFCAPS-TR-2014-0001

Cyber Test Form Development and Follow-on Cyber Applications

D. Matthew Trippe
Karen O. Moriarty
Adam S. Beatty
Tirso E. Diaz
Human Resourses Research Organization
66 Canal Center Plaza, Ste 700
Alexandria, VA 22314-1578
703.549.3611

Prepared under:
W911NF-11-D-0001, DO 0149
Battelle Memorial Institute
505 King Avenue
Columbus, OH 43201-2696

Prepared for:
Gregory G. Manley
AFPC/DSYX
550 C Street West
Randolph AFB, TX 78150

October 2014

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report was cleared for release by HQ AFPC/DSYX Strategic Research and Assessment Branch and is releasable to the Defense Technical Information Center.

The views expressed are those of the authors and not necessarily those of the United States Government, the United States Department of Defense, or the United States Air Force. In the interest of expediting publication of impartial statistical analysis of Air Force tests SRAB does not edit nor revise Contractor assessments appropriate to the private sector which do not apply within military context.

Federal Government agencies and their contractors registered with Defense Technical Information Center should direct request for copies of this report to:

Defense Technical Information Center - http://www.dtic.mil/

Available for public release. Distribution Unlimited. Please contact AFPC/DSYX Strategic Research and Assessment with any questions or concerns with the report.This paper has been reviewed by the Air Force Center for Applied Personnel Studies (AFCAPS) and is approved for publication. AFCAPS members include: Senior Editor Dr. Thomas Carretta AFMC 711 HPW/RHCI, Associate Editor Dr. Gregory Manley HQ AFPC/DSYX, Dr. Lisa Hughes AF/A1PF, Dr. Paul DiTullio AF/A1PF, Kenneth Schwartz HQ AFPC/DSYX, Johnny Weissmuller HQ AFPC/DSYX, Dr. Laura Barron HQ AFPC/DSYX, Dr. Mark Rose HQ AFPC/DSYX, and Brian Chasse HQ AFPC/DSYX.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
15-10-14	Final	Oct, 2013 – Oct 2014

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUM: W911NF-11-D-0001
Cyber Test Form Development and Follow-on Cyber Applications	5b. PROGRAM ELEMENT NUMBER: DO 0149
6. AUTHOR(S) D. Matthew Trippe, Karen O. Moriarty, Adam S. Beatty, Tirso E. Diaz	5c. PROJECT NUMBER: 2014 No. 041
	5d. TASK NUMBER
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resourses Research Organization 66 Canal Center Plaza, Ste 700 Alexandria, VA 22314-1578	8. PERFORMING ORGANIZATION REPORT NUMBER 2014 No. 041
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFPC/DSYX 550 C Street West, Suite 45 Randolph AFB, Texas 78150	10. MONITOR ACRONYM AFPC/DSYX
	11. MONITOR REPORT NUMBER AFCAPS-FR-2014-0001

12. DISTRIBUTION/AVAILABILITY STATEMENT

Available for public release. Distribution Unlimited.

13. SUPPLEMENTARY NOTES: Prepared under: W911NF-11-D-0001, DO 0149, Battelle Memorial Institute, 505 King Avenue, Columbus, OH 43201-2696

14. ABSTRACT *(Maximum 200 words)*:

The current project to develop the Cyber Test represents a continuation of the previous work on the Information-Communication Technology Literacy (ICTL) test. The ICTL test was designed to predict success in entry-level training in cyber-related military occupations. Previous research on the ICTL test showed it was – and is – successful in this regard (Russell & Sellman, 2009, 2010; Trippe & Russell, 2011). The goals of this project are to (a) update and expand the Cyber Test item bank, (b) pilot test the items in an operational setting, and (c) develop additional parallel test forms.

15. SUBJECT TERMS

DoD Enlisted Cyber Test, Information-Communication Technology Literacy (ICTL) test

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			Gregory G. Manley
Unclassified	Unclassified	Unclassified	Unlimited	31	(210) 565-0130

Standard Form 298

CYBER TEST FORM DEVELOPMENT AND FOLLOW-ON CYBER APPLICATIONS

Table of Contents

List of Tables

List of Figures

CYBER TEST FORM DEVELOPMENT AND FOLLOW-ON CYBER APPLICATIONS

Introduction and Background

The Armed Services Vocational Aptitude Battery (ASVAB) is the cognitive test battery used by all US military services for selection and classification of enlisted trainees. At present, the ASVAB consists of nine subtests: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), Electronics Information (EI), and Assembling Objects (AO). The Armed Forces Qualification Test (AFQT) which combines the two verbal (WK, PC) and two math subtests (AR, MK) is used for selection for each of the Services. Numerous studies have shown that the ASVAB is a valid predictor of training and on-the-job performance (e.g., Campbell & Knapp, 2001; Ree & Earles, 1992; Welsh, Kucinkas, & Curran, 1990).

At the request of the Office of the Assistant Secretary of Defense, the Defense Manpower Data Center (DMDC) began a review of the ASVAB in 2005 because of concerns that the content was dated due to changes in the nature of military service (e.g., more diverse missions, more complex organizations and systems, and enhanced technology) that affect the nature of military work and the characteristics required of military personnel. An expert review panel was convened to consider the current status of the ASVAB program and to make recommendations for improvements and enhancements as well as implementing such changes. The panel met in December 2005 to review what was presented and to reach consensus regarding its recommendations. The review panel presented its findings (Drasgow, Embretson, Kyllonen, & Schmitt, 2006) in March 2006 which included 22 recommendations. One of the panel's recommendations was that research should be conducted to develop and evaluate a test of information and communications technology literacy.

In response to the ASVAB review, the Air Force Personnel Center (AFPC) initiated a project in October 2007 to develop and evaluate a test of information and communications technology literacy. Many military jobs involve information and communications technology. As a result, the ASVAB review panel speculated that an updated technical test along the lines of the ASVAB Electronic Information subtest might improve validity and classification efficiency. This recommendation was consistent with a 2006 report by the National Academy of Engineering and the National Research Council regarding technological literacy.

The current project to develop the Cyber Test represents a continuation of the previous work on the Information-Communication Technology Literacy (ICTL) test. The ICTL test was designed to predict success in entry-level training in cyber-related military occupations. Previous research on the ICTL test showed it was – and is – successful in this regard (Russell & Sellman, 2009, 2010; Trippe & Russell, 2011). The goals of this project are to (a) update and expand the Cyber Test item bank, (b) pilot test the items in an operational setting, and (c) develop additional parallel test forms.

Blueprint Validation

The test blueprint upon which the Cyber Test is based was originally developed in 2008 (see Russell & Sellman, 2009). The blueprint is organized hierarchically, with four broad content areas at the highest level. Subsumed within each broad content area are several sub-content areas that are more specific and focused. At the lowest, most specific level of the blueprint hierarchy are knowledge, skill and ability (KSA) statements that serve as the basis for item development.

Prior to developing new items for the Cyber Test, the blueprint underwent a content validity evaluation to determine the relevance of the original blueprint to contemporary entry-level training for cyber-related occupations. We administered a blueprint validation survey to 34 military subject matter experts (SMEs) in cyber-related occupations. The SMEs were service members from the Air Force (25), Navy (7), and Army (2). SMEs reported an average of 12 years of experience in cyber-related occupations, with a range between two and twenty-one years of experience. SMEs were provided the broad content areas, sub-content areas, and example/representative KSA statements from the original Cyber Test blueprint. They were asked to provide a rating of the appropriateness of the broad area, sub-content area, and example/representative KSA statements for an aptitude assessment designed to predict performance in entry-level training for cyber-related military occupations.

Table 1 summarizes the results of the SME blueprint survey. There was virtually no disagreement concerning the appropriateness of the first three broad content areas among the 34 SMEs. There was considerably less agreement on the appropriateness of the "Software Programming and Web Development" content area and its associated sub-content areas.

Table 1. Summary Responses to Whether Cyber Test Blueprint is Appropriate for Applicant Testing

Broad/Sub-Content Area	Agree	Undecided	Disagree
Networks and Telecommunications	34	0	0
Network Configuration & Maintenance	33	1	0
Telecommunications	25	8	1
Computer Operations	32	1	0
PC Configuration and Maintenance	33	0	1
Using IT Tools/Software	27	7	0
Security and Compliance	33	0	0
System Security	34	0	0
Offensive Methods	26	7	1
Software Programming & Web Development	22	7	3
Software Programming	23	6	5
Database Development & Administration	18	10	6
Web Development	17	12	5
Data Formats	20	9	4
Numbering Systems	19	10	5

Note. *n*=34

The need to address the disagreement observed over the Software Programming and Web Development content area as well as the need to gather more specific information necessary to guide item development efforts prompted a more extensive follow-up blueprint validation survey. A subset of the original 34 SMEs (8 Air Force, 6 Navy, and 2 Army) completed the more comprehensive follow up blueprint survey. These 16 SMEs reported an average of 9.4 years of experience, with a range between two and nineteen years of experience.

Within each broad content area, the SMEs were asked to indicate the five most important and five least important knowledge, skill and ability (KSA) statements for testing. They were also allowed to rate something as "Neither." KSA statements rated as "Most important" were given a score of five; statements rated as 'Neither' a score of three; statements rated 'Least important' a score of one. This rating format is not ideal from a psychometric perspective, but was chosen to make the task less burdensome for the respondents. Each respondent was asked to provide three ratings for each of 48 KSA statements (Importance, Needed at Entry, and Obsolescence) in addition to providing estimates of test content weights and judgments specific to the Software Programming and Web Development content area.

Table 2 displays the KSA statements that at least 10 of the 16 SMEs rated as "Most important" and Table 3 displays the KSA statements that at least 10 of the 16 SMEs rated as "Least important." The most important KSA statements found in Table 2 tend to be fairly general statements that concern fundamental concepts that often serve as the basis for higher level learning within each content category. Many of the least important KSA statements found in Table 3 relate to office productivity applications or hardware components.

Table 2. Most Important KSAs by Broad Content Category

Category	KSA Statement	M	SD	# Least	# Neither	# Most
NT	Knowledge of common network terminology	3.75	1.77	4	2	10
NT	Knowledge of network protocols and standards	4.25	1.44	2	2	12
NT	Knowledge of telecommunication protocols (e.g. TCP/IP, OSI (open systems interconnection) layers)	4.00	1.63	3	2	11
CO	Knowledge of basic computer concepts (bit, byte, CPU)	4.25	1.44	2	2	12
CO	Knowledge of how the operating system interacts with hardware, user processes, application and networked components	4.38	1.41	2	1	13
CO	Ability to search on-line and other resources to obtain information that will help solve a problem	4.50	1.37	2	0	14
SC	Knowledge of telecommunication protocols (e.g. TCP/IP, OSI layers (open systems integration), token rings)	4.25	1.24	1	4	11
SC	Knowledge of operating system vulnerabilities	3.88	1.63	3	3	10
SC	Knowledge of how worms and viruses work	4.13	1.26	1	5	10
SC	Knowledge of network vulnerabilities	4.50	1.15	1	2	13
SC	Knowledge of encryption and decryption methods	3.88	1.63	3	3	10
SPWD	Knowledge of basic language constructs (e.g., arrays, do-loops, if/then statements)	3.75	1.77	4	2	10
SPWD	Ability to write, modify, execute, and interpret simple scripts	3.88	1.63	3	3	10

Note. NT=Networking & Telecommunications; CO=Computer Operations; SC=Security & Compliance; SPWD=Software Programming & Web Development.

It is notable that the KSA statements related to office productivity applications rated as "Least important" were also rated as "Needed at Entry" (see next section). It may be that SMEs view these KSA statements as so basic and commonplace that they are not perceived as particularly important. Office productivity application knowledge and skill is becoming more common among high school students,

who are using them more in their high school work. It is possible that in the future the predictive efficacy of these KSAs may decrease.

The aforementioned results pertaining to the KSA statements can be used to exclude non-essential KSAs. For two reasons, however, some degree of caution is warranted in interpreting the results. First, the KSAs provided to the SMEs were originally identified as being both important and needed at entry for cyber-related occupations (Russell & Sellman, 2009). Because the KSAs were indicated as relevant to similar occupations in the past, we would seek relatively strong evidence to justify the exclusion of specific KSAs in this specific instance. Second, because of the relatively small SME sample size available, we were concerned about radical changes to the established blueprint. With these considerations in mind, we ultimately elected to exclude "Knowledge of biometric technologies" and "Knowledge of cabling and wiring installation procedures."

Table 3. Least Important KSAs by Broad Category

Category	KSA Statement	M	SD	# Least	# Neither	# Most
NT	Knowledge of cabling and wiring installation procedures	2.00	1.63	11	2	3
NT	Knowledge of telecommunication topologies (e.g., entry points, exit points, network mapping)	2.38	1.89	10	1	5
CO	Ability to connect PC hardware components (e.g., monitor, printer)	2.25	1.77	10	2	4
CO	Knowledge of word processing software	1.75	1.24	11	4	1
CO	Knowledge of spreadsheet software	2.13	1.63	10	3	3
CO	Knowledge of presentation software	2.25	1.77	10	2	4
SC	Knowledge of biometric technologies	1.75	1.44	12	2	2

Note. NT=Networking & Telecommunications; CO=Computer Operations; SC=Security & Compliance; SPWD=Software Programming & Web Development.

In addition to Importance ratings, the SMEs were asked to provide "Needed at Entry" ratings for each KSA statement. Specifically, SMEs were instructed to indicate whether each KSA should be acquired prior to enlistment (needed at entry) or following enlistment (not needed at entry). Table 4 contains KSA statements indicated by at least half of the SMEs as being needed at entry. There was one KSA with complete agreement as to whether it should be acquired prior to enlistment – "Knowledge of the functions and operation of typical PC hardware and peripherals" – from the Computer Operations content category. Of note here are the office productivity KSAs (e.g., word processor, spreadsheet), which were not considered to be important, on average. We think the reason more KSAs are not endorsed as needed at entry is because they are perceived by SMEs as more advanced than they actually are.

Table 4. KSAs Where at Least Half of the Respondents Endorsed as Needed at Entry

Category	KSA
NT	Knowledge of common network terminology
NT	Knowledge of the purpose and functions of network hardware (e.g., routers, switches)
NT	Knowledge of network essentials (e.g., hub v. switch; types of networks)
NT	Knowledge of common network terminology
NT	Knowledge of the purpose and functions of network hardware (e.g., routers, switches)
CO	Knowledge of basic computer concepts (bit, byte, CPU)
CO	Knowledge of the functions and operation of typical PC hardware and peripherals
CO	Knowledge of different types of memory (e.g., ROM, RAM)
CO	Ability to connect PC hardware components (e.g., monitor, printer)
CO	Knowledge of word processing software
CO	Knowledge of spreadsheet software
CO	Ability to search on-line and other resources to obtain information that will help solve a problem
SC	Knowledge of telecommunication protocols (e.g. TCP/IP, OSI layers (open systems integration), token rings)
SC	Ability to write at appropriate level for reader
SC	Knowledge of how worms and viruses work

Note. NT=Networking & Telecommunications; CO=Computer Operations; SC=Security & Compliance; SPWD=Software Programming & Web Development

The SMEs were asked to estimate the rate of obsolescence for each KSA statement using the scale shown in Table 5. Higher scores indicate slower rate of obsolescence.

Table 5. Obsolescence Rating Scale

Rating	Assigned Score
Likely to change in 6 months or less	1
Likely to change in 6 months to 2 years	2
Likely to change in 2 to 5 years	3
Likely to change in 5 to 10 years	4
Likely to change in 10 years or more	5
Not likely to change at all	None[a]

[a]The "Not likely to change at all" rating was not part of the original rating scale developed in 2008 and was not assigned a score so current results could be compared to those obtained in 2008.

Because we used the same scale in the original SME online survey (Russell & Sellman, 2009), Table 6 includes the means from both the current and original SMEs. The current results suggest a faster rate of obsolescence for Computer Operations content than the original results, but slightly slower rate for the other three categories.

Table 6. Obsolescence Ratings by Category

Category	Current Mean	Original Mean
Networking & Telecommunications	3.49	3.21
Computer Operations	2.81	3.60
Security & Compliance	2.94	2.39
Software Programming & Web Dev	3.90	3.84

Note. Lower scores indicate a faster rate of obsolescence.

SME ratings of the importance, stability and "needed at entry" status of each KSA statement were used to guide item development tasks described in the next section. Item writers were instructed to focus their efforts on KSA statements that were identified as most important, highly stable and needed at entry.

SMEs were asked to estimate the test content area weights using multiples of five percentage points. Results are presented in Table 7. There was considerable variability in the weight estimates, likely reflecting the different occupational perspectives of the SMEs. The current mean estimated weights differ slightly from the original weights derived in 2008, which were 30%, 35%, 25%, and 15% for Networking & Telecommunications, Computer Operations, Security & Compliance, and Software Programming & Web Development, respectively. The operational test forms ultimately developed in 2011 contained content weights of 25%, 40%, 25%, and 10% percent. Deviations from the SME mean weights reflect the constraints of available items in form assembly. We used weights of 35%, 35%, 20%, and 10% to guide item development efforts described in the next section. This weighting scheme represents a compromise between the original and current SME weight estimates.

Table 7. Category Weights

Category	M	SD	Min	Max
Networking & Telecommunications	35.67	15.34	10.00	75.00
Computer Operations	29.33	14.62	5.00	70.00
Security & Compliance	21.33	4.81	10.00	30.00
Software Programming & Web Dev	13.67	7.90	0.00	20.00

Item Development

We recruited information technology (IT) experts to serve as item writers. We contacted faculty and posted fliers at local universities and technical schools, posted an advertisement on a popular internet forum, and contacted a local IT consulting firm to identify IT experts with sufficient time and interest to assist with item development. We invited those interested to submit their resumes, and those who seemed a good fit with the requirements of the project were then asked to develop 3-4 items with some basic guidance as a 'try-out.' The purpose of the try-out was to allow us to evaluate their item-writing potential and to allow potential item developers to assess the amount of work involved in

item development. In the end, we hired three IT experts local to HumRRO's Louisville, KY office and one from out-of-state1.

Cyber Test item developers underwent 5.5 hours of training in item development in early October 2012. The training reviewed a number of important aspects to developing quality items, including the purpose of the test, the demographics of the target population, and best practices in test item development. We also spent time reviewing current or recently-retired ICTL items along with their item statistics (e.g., difficulty and discrimination) and HumRRO's Guidelines for Sensitivity and Bias Review (Waters, 2008).

Even with such training, item review is still necessary to help mitigate construct-irrelevant factors' effects on test reliability and validity. Item review is typically an iterative process involving many steps and people. It confirms the items are (a) content valid, (b) appropriate for the test's purpose, (c) appropriate for the target population, (d) current in their content, and (d) correctly keyed. Each of the 200 newly-developed Cyber Test items underwent three levels of review – editorial, technical, and military.

After development, items immediately underwent an editorial review. Those with significant edits or comments were returned to the author for revision; a second editorial review was performed after edits were made. The editorial review was primarily concerned with grammar, reading level, appropriateness for the test and population, and adherence to HumRRO's Guidelines for Sensitivity and Bias Review (Waters, 2008).

The technical and military reviews were both concerned with whether the content was current and the key was correct, although the military reviewers provided an additional check on the appropriateness of the items to the target population. Thus, after the editorial review was completed, each item underwent a technical review performed by two item writers, neither of whom were the item's author. Their feedback was provided to the author who incorporated or otherwise addressed the edits (i.e., authors were not required to implement edits with which they disagreed). Another editorial review was performed on the items following the technical review if edits were made. Seven military cyber experts performed the military review. We had three experts from the Air Force, three from the Navy, and one from the Army. One hundred-sixty items were reviewed by two military cyber experts with the remaining 40 reviewed by one. Their feedback was incorporated or otherwise addressed, and the items were finalized.

Item Review

Editorial Review

The item writers developed 10 to 20 items per week and submitted them for an editorial review during the months of October and November. Early in development, many items had to be re-written to target content appropriate for our population or the test's purpose. This was less of an issue as the item writers gained more experience.

Two items were found to belong to different test content categories than those originally indicated. No items were found to be in violation of sensitivity guidelines regarding (a) offensive or exclusionary language, (b) stereotypes, (c) ethnocentrism. There were many instances where syntax and

[1] Several weeks into the project one of our local item writers had to discontinue work, but we were able to develop 200 items with the remaining three item writers

vocabulary had to be simplified when the same notion could be conveyed without introducing unnecessary verbal load.

Information technology (IT) is a topic that is differentially familiar to certain groups of test takers, which is commonly referred to as the 'digital divide.' A 2011 Washington Post article reported that 32% of U.S. households do not have Internet access at home. This will of course, disadvantage some test takers, but is unavoidable considering the purpose of the test. A large portion of the test covers Computer Operations, which includes some topics with which many test takers may have familiarity through school (e.g., word processing software).

Technical Review

Part of the item writing training included how to review test items. When reviewing each other's items, the item writers were asked to address the following questions and make specific suggestions:

- Is the item in the correct content area, or would it be better suited in another one? If so, which one?
- Is the item based on trivial or obscure knowledge?
- Is the content current?
- Is the item appropriate for our target population?
- Is the stem understandable? Does it need to be reworded? If so, how?
- Are the distractors plausible? If not, how can we change them?
- Is the key correct? Is there only one correct response? If not, how can we correct it?

As a result of the technical review, a small number of items underwent major revision, and about one-third underwent minor to moderate revision. These revisions primarily concerned the correctness of the key and whether there was only one correct response, plausibility of the distractors, and the currency of the item's content.

Military Review

The military review was handled via email. All reviewers signed a non-disclosure agreement, which set out rules for saving items while under review and destroying them upon completion of the review. SMEs reviewed between 40 and 60 items each across the different content categories during late November/early December 2012. As previously noted, 160 of the items were reviewed by two SMEs and the remaining 40 were reviewed by one. Of the seven cyber experts, six provided ratings while the seventh answered questions and helped address discrepancies in reviews on an ad hoc basis.

The cyber experts were provided a project brief that explained the purpose of the test, the target population, organization of the test, and their task. They were also given a spreadsheet with the identification numbers (IDs) of the items they were reviewing and a listing of the KSAs by test content category. They were instructed to enter their ratings and feedback into the spreadsheet and return to HumRRO.

SMEs were asked to make two judgments and offer any suggestions or comments they felt necessary. Specifically, they were asked to indicate whether the key was correct (yes/no) and to make a judgment on the item's quality. The Quality rating was made on the following Likert scale:

1 = Extremely low quality
2 = Low quality
3 = Neither low nor high quality
4 = High quality
5 = Extremely high quality

To make the quality rating, the SMEs were asked to consider how well the item measured the content category, the currency of the content, the appropriateness for the target population, and their general reaction to the item.

The SMEs provided very helpful feedback, and based on their feedback, minor edits were made to 12 items and 10 items were dropped. It was often the case that the cyber experts did not agree with each other on the quality rating. For example, one rater gave most of the Computer Operations items he reviewed a 1 or 2 because he felt they were not "truly" cyber, whereas the other rater gave them 4s or 5s. Another rater disapproved of a reference for a few items and rated them 1s or 2s, whereas the other rater rated these items highly. This is confirmed in the rater reliability estimates where we obtained an ICC(1,k) of .056 and G(q,k) of .1282. Only the 160 items with two ratings were included in the reliability analyses. Tables 8 and 9 summarize the quality ratings, and Figure 1 provides a histogram of the item mean ratings. Figure 1 includes the 40 items with only one rating where a mean is actually a single rating.

Table 8. Summary Quality Ratings by Content Category

Category	M	SD
Overall	3.56	0.74
Networking & Telecommunications	3.60	0.52
Security & Compliance	3.49	0.68
Computer Operations	3.54	0.96
Software Programming & Web Dev	3.67	0.62

Table 9. Summary Quality Ratings by Rater

Rater	M	SD
1	3.49	0.79
2	3.32	1.01
3	4.17	0.57
4	3.68	0.77
5	2.94	1.18
6	3.54	0.59

Note. There were a total of 7 cyber expert reviewers. Six provided ratings; the seventh helped to clarify disagreements, but did not provide ratings.

[2] G(*q*,*k*) is an interrater reliability estimate especially suited to ill-structured data such as these where six different raters provided multiple ratings on 160 items. Unlike ICC(1,*k*), it allows researchers to distinguish between two sources of error – rater main effects and item x rater interaction effects and residual error. See Putka, Le, McCloy, & Diaz (2008) for a detailed discussion.

The military experts' high level of disagreement is not different from what occurs with standard item reviews. Each reviewer has his or her own biases and perspective, which is why it is important to include multiple rounds of reviews with different people. Military SMEs are better able to judge the applicability of the items to the target population than civilian experts who do not have much experience in this regard.

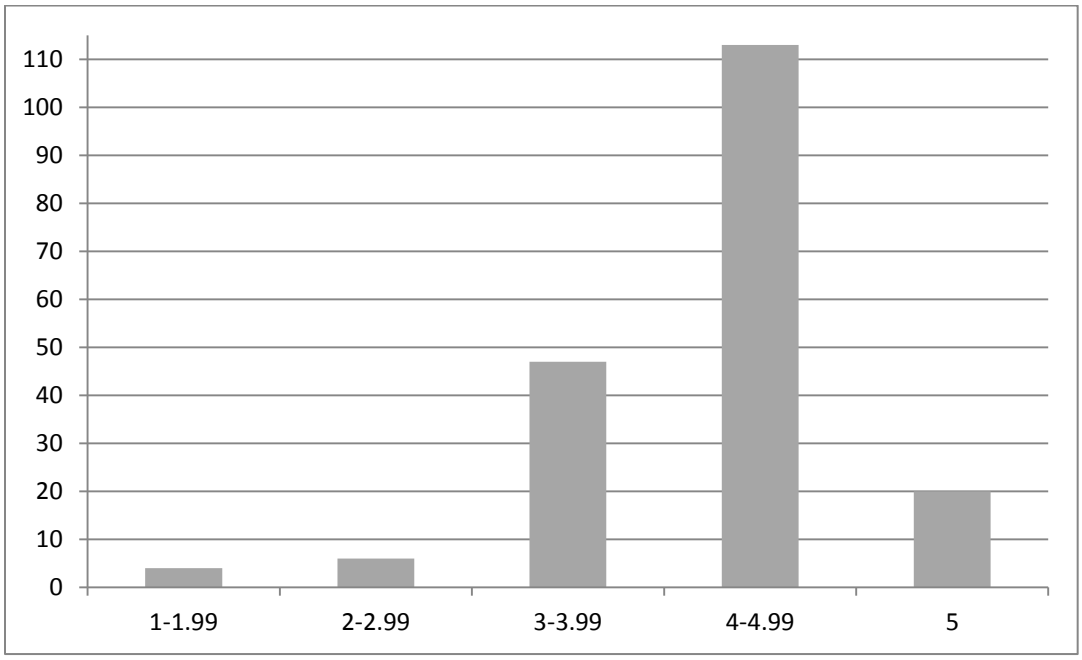

Figure 1. Frequency distribution of mean item ratings.

All 200 newly-developed Cyber Test items underwent multiple levels of review. As a result nearly all were edited to some extent, and 10 were dropped. One hundred-ninety items remained for pilot testing.

Pilot Test

Item Administration

The 190 experimental items developed in the previously described sections were provided to the Defense Manpower Data Center (DMDC) for administration on the ASVAB platform. Experimental items were "seeded" within existing Cyber Test forms in a manner similar to that of experimental ASVAB items. More specifically, 10 randomly selected experimental items were administered to 31,382 Service applicants between September of 2013 and June of 2014 such that each applicant received a different combination of 10 items randomly selected from the set of 190 items. Each experimental item was administered to an average of approximately 1,500 applicants. This kind of randomization effectively controls for many potential extraneous factors (e.g., order effects) encountered in traditional pilot testing.

The sample used to conduct all calibration and equating analyses was created by limiting the full data set in several ways. First, we only included applicants who were testing for the first time (i.e., eliminating retests or confirmation tests; *n* =3,001). We also eliminated a small number of applicants who spent less than three minutes taking the assessment. Three minutes is a liberal criterion used to remove only the most extreme outliers and is merely one of the data screens applied. Finally, we removed data for anyone who scored at or below chance on the existing Cyber Test and also scored above average on the AFQT. We chose this screening tactic rather than simply removing individuals who scored at or below chance because we expected to see the full range of aptitude in such a large sample. That is, the point of such a data screen is to remove individuals who we suspect lack motivation for taking the test. Nevertheless, in a sample of over 31,000 there will be a small number of individuals who, by virtue of exerting cognitive effort on the assessment, will score at or below chance levels. The compound rule was devised to remove individuals who showed other indications of a reasonably high level of aptitude, but performed very poorly on the Cyber Test. Data screens related to potentially unmotivated applicants combined resulted in the removal of only 83 individuals, which is likely a reflection of the fact that the Cyber Test was presented seamlessly with other pre-enlistment assessments. That is, we expect a fairly high level of motivation in an applicant sample where examinees do not know the difference between operational and experimental assessments or operational vs. experimental items. Characteristics of the sample used for calibration and equating analyses are summarized in Table 10.

Table 10. Demographic characteristics of the calibration sample

Characteristic	*n*	% of Sample
Service/Component		
Army Guard	165	0.58
Army Regular	252	0.89
Army Reserve	46	0.16
Air Force Guard	1,901	6.72
Air Force Regular	7,877	27.84
Air Force Reserve	1,204	4.25
Marine Regular	91	0.32
Marine Reserve	13	0.05
Navy Regular	15,853	56.02
Navy Reserve	852	3.01
Coast Guard Regular	24	0.08
Coast Guard Reserve	1	0.00
Other	19	0.07
Gender		
Female	7,365	26.03
Male	20,908	73.89
Unknown	25	0.09
Race		
American Indian	353	1.25
Asian	1,217	4.30
African American	5,480	19.37
Caucasian/white	18,648	65.90
Hawaiian/Pacific	264	0.93
Other	1,370	4.84
Decline to Respond	966	3.41
Ethnicity		
Hispanic or Latino	4,147	14.65
Not Hispanic or Latino	21,999	77.74
Decline to Respond	2,152	7.61
Total	28,298	100.00

Item Analysis

All Cyber Test items were analyzed using an Item Response Theory (IRT) measurement model known as the Three Parameter Logistic Model (3PL) (Lord, 1980; Lord & Novick, 1968). In essence, IRT assumes that test item responses by examinees are the result of underlying levels of ability possessed by those individuals. IRT provides a seamless approach to a variety of test analysis, development, and reporting activities. IRT is facilitated by fitting, or calibrating, statistical models to examinee responses. Application of these statistical models results in the simultaneous scaling of item difficulty and examinee (population) ability. Calibration was executed via the software program MULTILOG (Thissen, 2003).

IRT algorithms search for "item parameters," which capture a nonlinear relationship between ability and the likelihood of correctly answering each item. In the 3PL model, the probability that an examinee with an ability estimate ϑ responds correctly to item *i* is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp\left[-1.7a_i\left(\theta - b_i\right)\right]}$$

where a_i is the item discrimination, b_i is the item difficulty and c_i is the pseudo-guessing parameter.

Items that fit the IRT model will exhibit a pattern of lower probabilities of correct responses from low-ability applicants and higher probabilities of correct responses from high-ability examinees. This is reflected in an item characteristic curve (ICC) as depicted in Figure 2.

Items vary in difficulty such that the position of the point of inflection on the ICC is higher or lower (i.e., to the right or to the left) along the ability (theta) scale. For example, the point of inflection of the curve for the sample item in Figure 2 is centered at zero, the mean on the ability scale. An efficient test will be composed of items with ICCs similar to that depicted, but with varying difficulties ("B" parameter) that discriminate along the entire ability scale, which is typically called "theta." Item characteristic curves also differ in their lower asymptotes (related to how easy it is to get the item correct by guessing, or the "C" parameter) and the gradient of their slopes at the inflection point (i.e., "A" parameter).

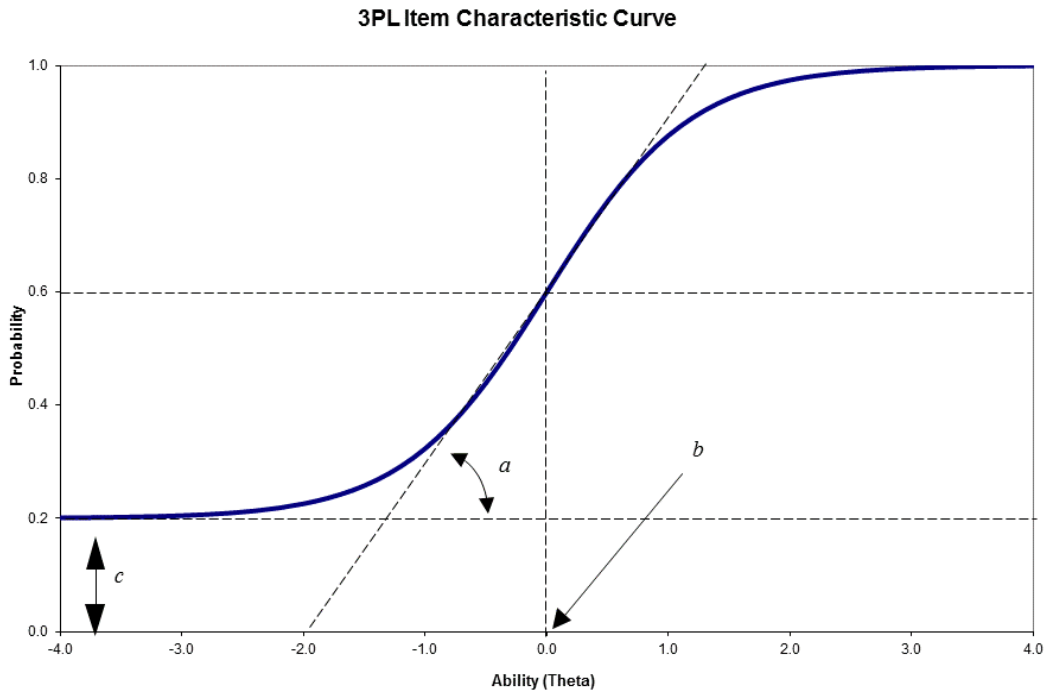

Figure 2. Example item characteristic curve in the 3 parameter logistic model.

Parameter Calibration & Equating

Each individual Service applicant in the calibration sample was administered one of two 29 item "operational" Cyber Test forms (see Trippe & Russell, 2011) and 10 randomly seeded experimental items. Each of the 190 experimental items was administered to an average of 1,489 individuals in the randomized design. IRT parameters are traditionally calibrated in a Marginal Maximum Likelihood (MML) procedure in which algorithms search for parameter values as well as ability values in an iterative fashion. We initially calibrated parameters for the combined item pool (i.e., operational and experimental) in the MML framework by progressively trimming the experimental items from inclusion in the calibration based on results of the IRT calibration and classical test theory (CTT) indices of item

quality (p-values and item-total correlations). That is, the first calibration included all operational and all experimental items and subsequent calibration attempts removed experimental items with poorly estimated parameter values (i.e., extreme or out of bounds values) or undesirable CTT statistics (e.g., low or negative item-total correlations). Although it was generally the case that experimental items with poor parameter estimates tended to be those with undesirable classical statistics we observed several exceptions to this convention. HumRRO has previously observed that the item-total correlation can under-represent non-linear functioning of difficult items and this functioning is often appropriately captured by the IRT model.

Parameter calibration in the traditional MML framework proved to be somewhat unstable and difficult to manage given the characteristics of the experimental item pool, which includes several relatively difficult items as well as a relatively high ratio of experimental to operational items. Item parameter values derived in MML calibration can become "contaminated" by other poorly calibrated items in the pool because of the joint estimation of parameter and ability values. This difficulty with traditional MML calibration led us to explore a "maximum likelihood for fixed theta" approach whereby parameter values are derived from a fixed or "known" ability value and an array of item responses. In this approach, we calibrated item parameters for the 58 operational items in the traditional MML framework. We then scored each of the applicants in the calibration sample using these operational parameter values alone. The 28,298 theta estimates were then standardized to a distribution with a mean of zero and standard deviation of one to counteract the "compression" that often results from maximum a posteriori (MAP) scoring in IRT. Parameter estimates for the 190 experimental items were then calibrated in the fixed theta framework. The fixed theta framework has a few advantages related to the stability of the calibration. Individual item parameter values are derived independently such that an item with poor parameter estimates cannot influence any other item's parameter estimates. The fixed theta calibration also strongly ties the parameter estimates to the original operational construct, which minimizes the influence of potential construct drift that can result from an off-topic or otherwise poorly functioning experimental item.

Item parameter estimates calibrated in the analyses just described were on a somewhat arbitrary scale that needs to be linked back to the original operational scale established in 2011 by an equating process. This process involves using the operational items administered in both 2011 and for this effort as "anchor items." We applied the Stocking-Lord (1983) procedure to establish a common scale. The Stocking-Lord procedure uses item parameters from the current effort and the original 2011 calibration to calculate test characteristic curves (TCCs) for each set of parameters. A transformation multiplier and additive constant (M1 and M2) are then calculated to transform the current TCC to match the original TCC as closely as possible.

Operational item parameters that served as anchor items in this procedure were evaluated for potential parameter drift from 2011. Anchor parameters were placed on a common scale. Then, values of the squared differences were calculated at 31 quadrature points (the same used in the Stocking/Lord procedure) and the mean of the 31 squared differences was computed for each item. Items were flagged if their mean squared difference (or mean d-square) was greater than expected, compared to an empirically derived sampling distribution of squared difference values. Nine of the 58 anchor items demonstrated statistical evidence of parameter drift. Three of the nine items demonstrated severe drift (i.e., exceeded the 99th percentile of the empirically derived sampling distribution of squared difference values) and were eliminated from equating on that basis alone. The remaining six items demonstrated less severe statistical evidence of drift and were evaluated for potential construct irrelevant variance to explain the drift. It was determined that one of the remaining six items was potentially obsolete and was also eliminated from the equating process. The Stocking-Lord (1983) procedure was then implemented using a total of 54 operational items as anchors. The resulting constants (M1 = 0.85641, M2 = 0.13765) were then used to transform all parameters to the original operational scale.

Post Hoc Sensitivity Review

A subset of the 190 experimental items underwent a *post hoc* sensitivity review based on statistical evidence of differential item functioning (DIF). Although the total calibration sample available was relatively large, only about five percent of the sample responded to any individual item. Analyses based on dividing the sample into subgroups were therefore limited and less than ideal. Nevertheless, we conducted analyses in five subgroup samples: males, females, non-Hispanic Blacks, non-Hispanic Whites and Hispanic Whites. These groups were chosen to be consistent with designations used by the ASVAB testing program (Defense Manpower Data Center, 2014). IRT-based DIF analyses were conducted for three comparisons: male vs. female, non-Hispanic White vs. non-Hispanic Black and non-Hispanic White vs. Hispanic White. In each comparison, a subgroup specific fixed theta parameter estimate was calibrated and the area between ICCs was computed. Items that were flagged as exceeding the 95th percentile of an empirical distribution of ICC gap measures were subject to sensitivity review.

This approach to DIF is admittedly weak with respect to sufficient subgroup sample sizes at the experimental item level and almost certainly resulted in an exceedingly high rate of false positive identification. Nevertheless, the fixed theta calibration did allow us to capitalize on the full available sample size to define the ability distribution in each subgroup (female n = 7,365, male n = 20,908, non-Hispanic Black n = 4,899, non-Hispanic White n = 14,183 and Hispanic White n = 3,496). Moreover, we did not conclude that an item was biased based on statistical evidence alone. Differences in relative difficulty may in fact represent construct relevant variance that is not necessarily bias. That is, there may be true differences in the construct across groups. An item or test cannot be said to be truly biased unless the source of the differential functioning is determined to be construct irrelevant. This requires logical analysis of the item or test content. The classic example is mathematical reasoning items presented in the context of batting averages. These items tend to exhibit differential functioning such that females of equal mathematical reasoning ability are less likely to get the item correct. Such items can be said to be biased because, although they assess the intended construct of mathematical reasoning, they also measure a second construct irrelevant dimension of baseball knowledge (Camilli & Shepard, 1994).

We therefore cautiously proceeded with the DIF analyses, fully aware of the relatively large degree of sampling error that would influence results. A total of 39 otherwise viable (i.e., deemed

acceptable in the *post hoc* item review described later) items were identified as demonstrating statistical DIF in one or more of the subgroup comparisons.

To evaluate the flagged items, we met with two information technology (IT) SMEs. One SME is a non-Hispanic Black male the other a non-Hispanic White female. Both SMEs had previously served as item writers in 2008. We presented background information on the ICTL/Cyber Test project as well as an introduction to measurement bias and the concept of construct irrelevant variance. The group also reviewed 6 sensitivity guidelines based on sensitivity review procedures we have used in the past (Russell & Sellman, 2008b).Those guidelines are summarized in Table 11.

Table 11. Summary of Sensitivity Review Guidelines

Guideline	Topic
1	Avoid the use of stereotypes.
2	Avoid ethnocentrism.
3	Do not use language or topics that may be differentially familiar to certain groups of test takers.
4	Do not use language that is exclusionary, offensive or unfamiliar to certain groups.
5	Avoid using difficult words, figures of speech, idioms, or complex syntactic structures that are not required to assess the construct being measured.
6	Illustrations, graphics, and other visual stimuli should be required to measure the intended construct and should depict different groups equally, in a wide range of societal roles and contexts.

We discussed each of the guidelines in more detail and gave examples of each. After covering the instructional material, we presented and discussed the flagged items one-by-one. Generally speaking, items did not necessarily favor one group vs. another in a systematic way. SMEs did notice that racial and ethnic minority groups tended to underperform on items related to using office productivity tools (e.g., Microsoft Excel or Word) and hypothesized that this trend was related to access or socioeconomic status rather than race or ethnicity. Despite these general trends, the sensitivity review group found no content-based evidence of bias in any of the items. The conclusion in each case was that the statistical evidence of DIF was due to either sampling error (which was admittedly large in these analyses) or true differences between subgroups in particular content areas, but not bias. All items flagged for DIF were retained in the item pool.

Post Hoc Item Quality Review

The ultimate goal of this project was to assemble additional parallel Cyber Test forms as an extension of the 29-item, non-overlapping forms developed in 2011. We decided to use the items on the existing Cyber Test forms and the newly developed experimental items as the basis for form assembly. That is, instead of building additional forms parallel to the two existing forms, we combined the 58 items from the existing forms with the 190 newly developed into a single pool from which to assemble new parallel forms. The primary driver of this decision was the opportunity to review the items in the original forms for potential obsolescence and create forms with entirely current content. Three psychometric SMEs with working knowledge of the content areas covered in the test blueprint independently reviewed the 58 operational Cyber Test items and flagged any item suspected to be subject to potential obsolescence. The three SMEs then met to discuss and resolve any disagreements among the obsolescence ratings. After consulting with a fourth IT SME regarding a small number of technical items,

consensus was reached that nine items were potentially obsolete. In most cases, the items were still relevant and functioning as intended but would likely become less effective over time because of references to outdated software versions (e.g., Windows Vista, XP) or technology concepts that were "common" at the time the item was written but are now less common (e.g., wired connection of peripheral devices). That is, CTT and IRT based indices of item quality suggested that most of these items were still of high quality, but concerns over the content were the primary driver of the decision to remove them from the pool.

As described in the section on item development, great care was taken to ensure the quality of the content in the 190 newly developed experimental items. Nevertheless, item quality must also be evaluated in terms of psychometric indicators. Three psychometric SMEs independently reviewed the content of each experimental item in the context of available psychometric indicators of item quality, which included (a) the p-value or proportion of applicants who endorsed the keyed response (b) the biserial item-total correlation (c) the proportion of examinees endorsing each distractor response (d) the distractor-total correlation (e) 3PL IRT item parameters and associated ICC and (f) an IRT-based information index[3]. As mentioned earlier in this report, HumRRO has previously observed that the item-total correlation can under-represent non-linear functioning of difficult items and this functioning is often appropriately captured by the IRT model. The information index was included to provide an additional perspective of item quality in light of a relatively high number of difficult items. In other words, items that provide a relatively high amount of information at a given level of ability (theta) may still be psychometrically useful.

The psychometric SMEs, who also have working knowledge of the item content, independently rated each experimental item as either an item to "keep" or "drop" from the final item pool. If the SME indicated the item should be dropped, he selected a reason from a drop down menu ("Content flaw," "Needs IT SME review," "Obsolete," or "Psychometric") and also provided an open ended explanation for the decision to drop. "Content flaws" included such things as two possible correct answers, which are often revealed by positive distractor-total correlations, or typographical errors. Items rated as "Needs IT SME review" were often highly technical in nature and showed some ambiguous psychometric properties. A small number of items were rated with this reason code and discussed with an IT SME to confirm content quality. Items rated as "Obsolete" were those that referred to content that had become dated since the item was written or were likely not to remain current in the foreseeable future. Items rated as "Psychometric" demonstrated poor statistical evidence of item quality such as (a) low or negative item-total correlation (b) extremely high or low p-value (c) extreme or out of bounds IRT parameters or (d) positive distractor correlation(s). The reason codes for dropping items from the pool were not necessarily mutually exclusive. It is often the case that content flaws are reflected in psychometric indices. Undesirable psychometric characteristics may also simply indicate that an item is inappropriate for the applicant population and are not necessarily indicative of poor item content.

After all ratings were completed independently, the SMEs met to discuss and resolve any discrepancies in the keep or drop decisions. SMEs talked through their rationale for keeping or dropping any item where there was not 100% agreement until a consensus was reached. In a small number of cases, a fourth IT SME was consulted to clarify technical issues related to item content. After all

[3] The IRT-based information index was calculated across 13 points along the ability distribution. At every ability point, the average information statistic across all items was calculated. Any individual item whose information statistic was greater than the average information at a given ability level was considered to have the potential to be psychometrically useful, regardless of more traditional indices, such as the p-value and item-total correlation. The IRT information index was the number of times across the 13 ability points that the item reached this threshold.

discrepancies were resolved, SMEs agreed that 118 (62%) of the experimental items were acceptable for the next step of form assembly. The most frequently used reason code for dropping an item was "Psychometric." Items assigned a reason code of "Content flaw" were often because of a distractor that could be plausibly correct. "Obsolete" items generally referred to software versions that will soon be replaced or products that no longer exist (e.g., FireWire).

Form Assembly

The form assembly process began with a pool of 167 items (49 from the original Cyber Test forms and 118 newly developed items) whose characteristics are summarized in Table 12. The original form development process in 2011 resulted in two 29-item forms with no overlap. The forms were balanced with respect to (a) item content, (b) difficulty, (c) discrimination, (d) reliability, and (e) keyed responses. We also needed to consider item "enemies" (i.e., items that assess identical or highly similar content) when making form assignments. Balancing difficulty, discrimination and reliability of forms is accomplished in an IRT framework by matching form test characteristic curves (TCCs) and test information functions (TIFs). TCCs are simply the summation of ICCs within a given form. TIFs provide a test level index of the degree of measurement precision for a given ability level. We took the same general approach in the current form assembly effort, but slightly relaxed the constraint of non-overlapping items for reasons detailed below. We also generated two different form assembly solutions using the same item pool: (a) one solution comprising a smaller number (4) of relatively longer (40 items) forms and (b) one solution comprising a larger number (5) of relatively shorter (30 item) forms.

Table 12. Summary of 3PL Item Parameters in the Final Item Pool

Parameter	M	SD	Min	Max
Discrimination (A)	0.73	0.40	0.15	2.46
Difficulty (B)	0.27	1.76	-4.21	4.07
Pseudo-Guessing (C)	0.21	0.11	0.03	0.58

Note. n = 167.

A number of the form objectives are not possible to maximize simultaneously (e.g., a way to maximize reliability might be to load as many similar questions as possible onto a form, but this would not allow us to balance content between the forms). Therefore, in order to determine the optimal assignment of items to forms to balance the competing test specifications, we utilized Automated Test Assembly (ATA; van der Linden, 2005). Although ATA can refer to a variety of different algorithms for test assembly, a common approach is to use binary/integer programming to reframe the problem as a mathematical optimization process. Specifically, an objective function is identified, which is the quantity that is to be minimized or maximized, and each of the test specifications is recast as a mathematical inequality on the set of possible solutions. In order to solve our specific problem, we used the basic ideas presented in van der Linden (2005) and Diao and van der Linden (2011), but developed our own implementation in SAS using PROC OPTMODEL. The objective function we minimized was an equally weighted sum of the distance between the TIFs and TCCs of the forms at five representative quadrature points. We also specified the content-area, item key, item enemy, and number of overlapping items targets as constraints on the solution set.

Many applications of ATA are in the educational testing domain, where massive item pools and sample sizes are the norm. Although we had a large sample size, the item pool was small relative to the number of items we planned on assigning to forms. This primary implication of a smaller item pool is that it is more difficult to achieve the form assembly targets. One way that we attempted to mitigate

this problem was by relaxing a previous requirement that no items be shared across forms. By sharing up to five items across forms, this allowed us to more closely match the TIFs and TCCs, while simultaneously increasing reliability by enabling longer forms, and allowing us to include highly discriminating items that would otherwise push the TIFs and TCCs apart in certain ability ranges. There are five common items in the four form solution and four common items in the five form solution.

Figures 3 and 4 present the overlaid test characteristic curves in the optimized four and five form solutions, respectively. TCCs within each solution are nearly indistinguishable, although there is a greater degree of separation in the five form solution. Closely matching TCCs across the forms indicates comparable difficulty and is often referred to as “pre-equating.”

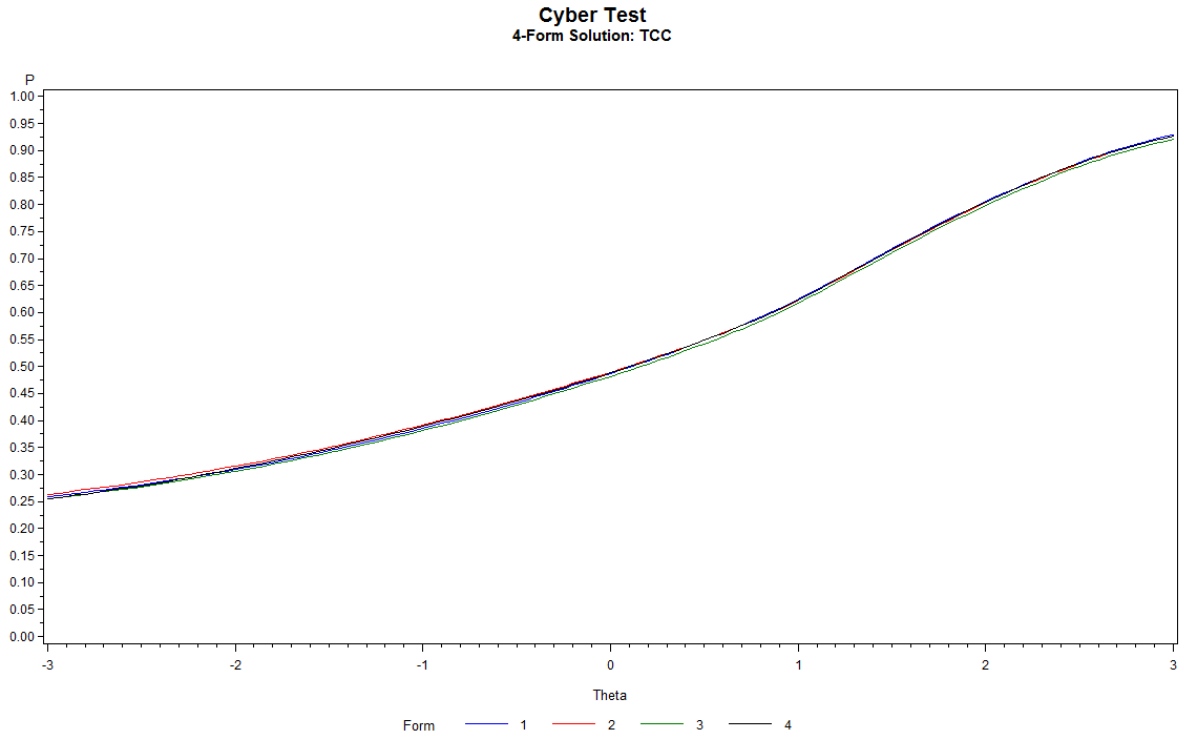

Figure 3. Overlaid test characteristic curves in the four form solution.

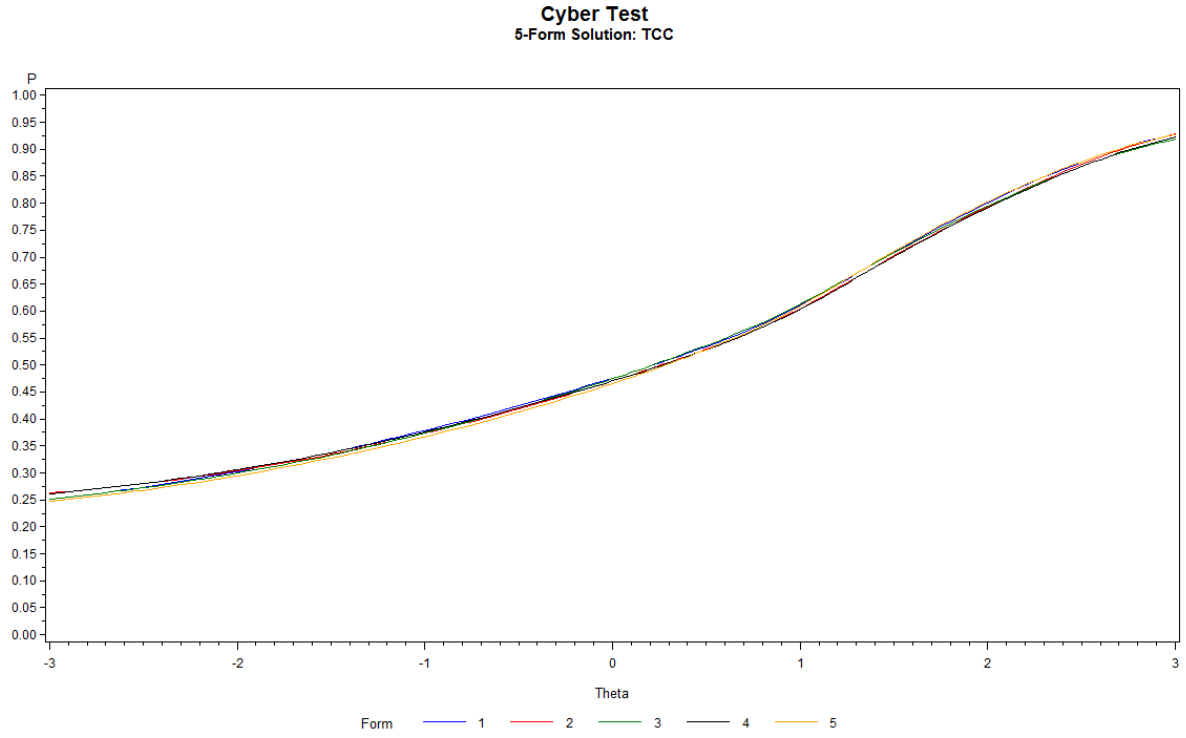

Figure 4. Overlaid test characteristic curves in the five form solution.

Table 13 contains the marginal reliability for each form in each solution. Marginal reliability is essentially an average reliability computed across the ability distribution. That is, a reliability estimate was computed for each level of theta and then the average of these reliability estimates weighted by the observed density of theta in the calibration sample was computed. The marginal reliability can be interpreted like a traditional index of reliability (e.g., Cronbach's alpha). Marginal reliability was computed over the entire range (-3 to 3) of ability as well as the higher end of the distribution (0 to 2.5) where the Cyber Test tends to function best. Nevertheless, distilling this information into a single index like the marginal reliability coefficient obscures the fact that IRT provides reliability estimates conditional on ability, which is fully captured in the TIFs seen in Figures 5 and 6. As expected, marginal reliabilities were larger for the 4-form solution, due to the longer forms (i.e., 40 items vs. 30 items).

Table 13. IRT Marginal Reliability by Form

Form	R_{xx} $-3 < \theta < 3$	R_{xx} $0 < \theta < 2.5$
4-Form Solution		
1	.77	.83
2	.76	.82
3	.77	.82
4	.76	.83
5-Form Solution		
1	.70	.77
2	.68	.76
3	.69	.78
4	.67	.76
5	.69	.78

The TIFs in figures 5 and 6 reveal a high degree of asymmetry in the information available in the Cyber Test item pool and in the derived forms. That is, the Cyber Test forms provide a relatively low degree of information or measurement precision at the low and even middle parts of the ability distribution, and a relatively high degree of information at the higher end (beginning just below a value of 1.0). These are fairly uncommon TIFs—as most tests tend to be most informative near the center of the ability distribution where most examinees are. The highly discriminating Cyber Test items (the discrimination parameter contributes most directly to information) tend to be more difficult items. The degree of asymmetry in information observed in Figures 5 and 6 would generally be considered undesirable for a test designed to provide the most information on the largest number of individuals. Nevertheless, the Cyber test is not currently being used to be diagnostic for low or even average ability applicants. The current cut score used by the Air Force is a value of 60 on the reporting scale, which is roughly equivalent to a value of 0.8 on the theta metric. The measurement precision provided by the Cyber Test forms is acceptable and even exceptional for the purpose of selecting and classifying high ability individuals into technical training.

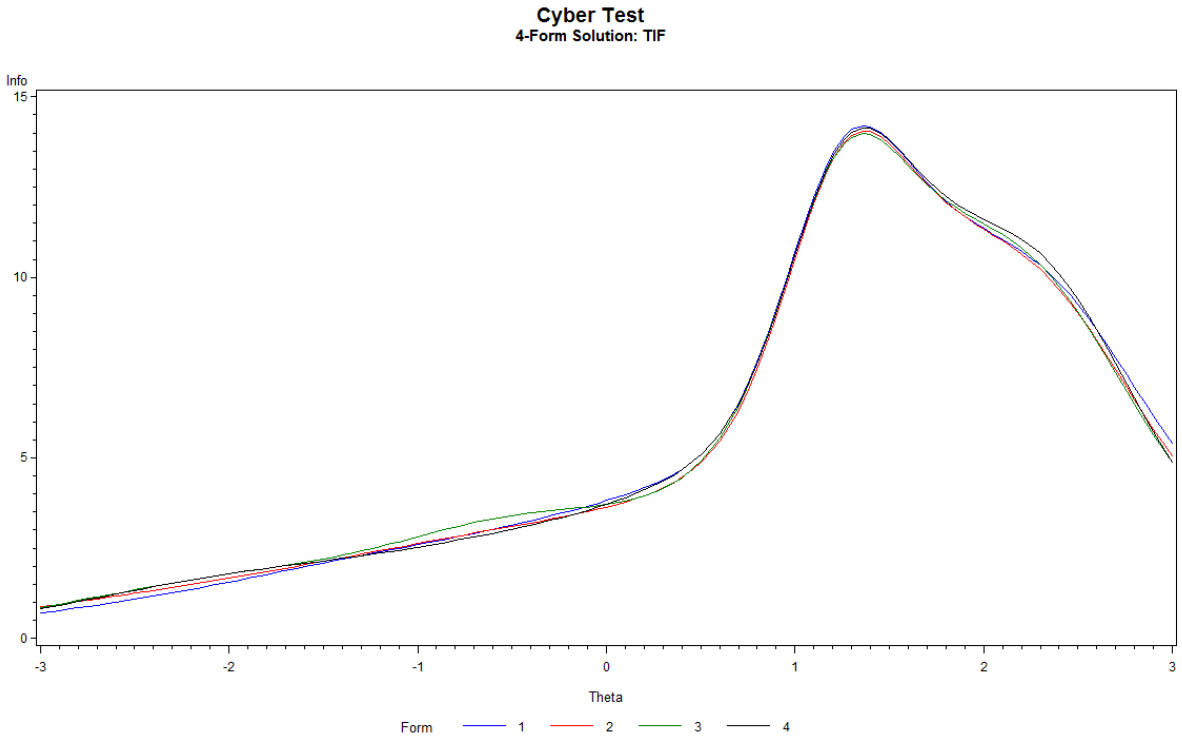

Figure 5. Overlaid test information functions in the four form solution.

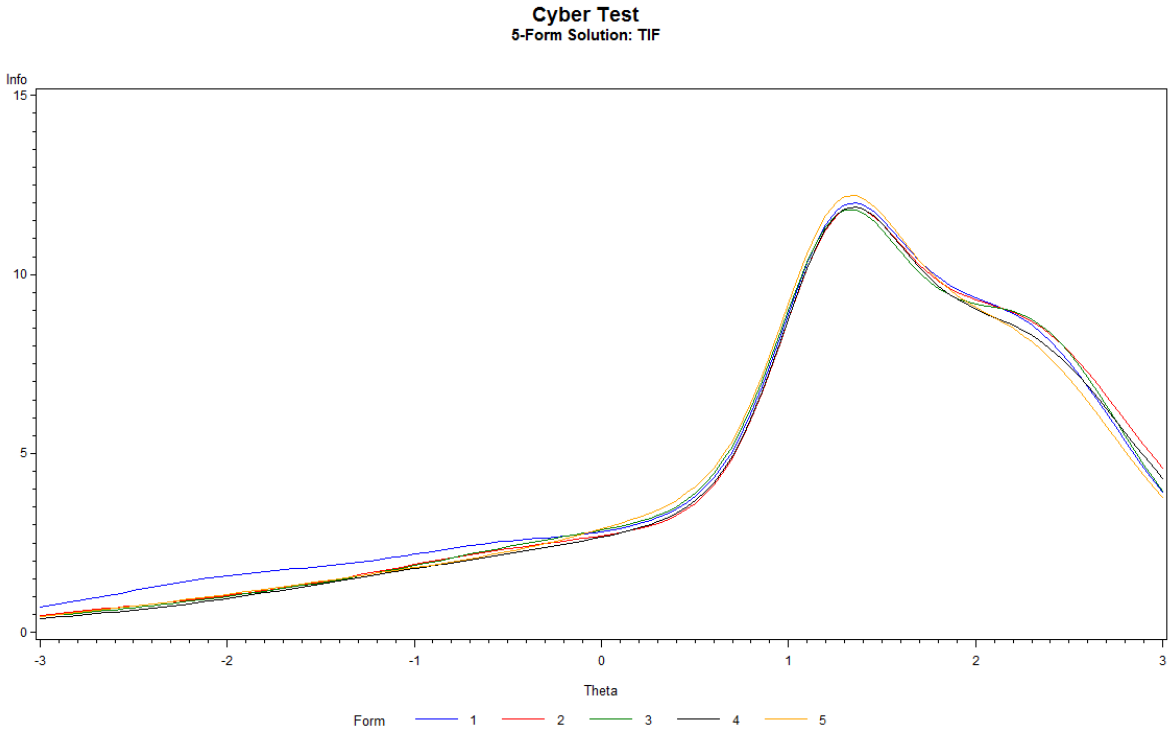

Figure 6. Overlaid test information functions in the five form solution.

Table 14 contains the content distribution for each form in the four and five form solutions. The ATA optimization algorithm began with a target distribution of 35%, 35%, 20%, and 10% for Networking & Telecommunications, Computer Operations, Security & Compliance and Software Programming & Web Development content areas, respectively. The target distribution was based on the blueprint validation work described above. The ATA algorithm was allowed a small amount of flexibility to deviate from the target distribution to satisfy other constraints in the optimized solution (e.g., TCCs, item enemies).

Table 14. Content distribution by form

Form	CO		NT		SC		SPWD	
	n	%	n	%	n	%	n	%
4 Form Solution. Length = 40 items								
1	15	37.5	13	32.5	8	20.0	4	10.0
2	16	40.0	12	30.0	9	22.5	3	7.5
3	16	40.0	14	35.0	6	15.0	4	10.0
4	16	40.0	12	30.0	9	22.5	3	7.5
5 Form Solution. Length = 30 items.								
1	11	36.7	11	36.7	6	20.0	2	6.7
2	11	36.7	11	36.7	6	20.0	2	6.7
3	11	36.7	12	40.0	5	16.7	2	6.7
4	11	36.7	11	36.7	6	20	2	6.7
5	12	40.0	11	36.7	5	16.7	2	6.7

Note. NT=Networking & Telecommunications; CO=Computer Operations; SC=Security & Compliance; SPWD=Software Programming & Web Development.

Summary and Conclusion

We developed the four and five form solutions from a common item pool to provide the Air Force with some flexibility in the implementation of the Cyber Test. Generally speaking, the forms in the four form solution are of higher quality with respect to indices of parallelism and individual quality. Forms in the five form solution contain fewer items and thus require less administration time. Having a fifth form also reduces item exposure to a greater degree and allows for additional retesting options. Nevertheless, we believe the benefits of the four form solution outweigh those in the five form solution and recommend its implementation at Military Entrance Processing Stations (MEPS) as the preferable option. Trippe and Russell (2011) found the mean assessment time for experimental 40-item ICTL test forms administered to over 50,000 applicants at MEPS to be approximately 12 minutes, with a range between 5 and 24 minutes. We expect assessment time to be comparable in the four form solution developed in this project. If the increased testing time associated with test forms that are longer than what is currently in use will be unacceptably disruptive, the five form solution is a viable alternative that is highly comparable to the two forms developed in 2011.

The forms in each solution have been pre-equated and every effort was taken to ensure equivalence of content, reliability, difficulty, and discrimination. Nevertheless, the entire form assembly process may capitalize on chance characteristics of the applicant sample used for development. Moreover, true parallelism of forms that contain non-identical item sets is an abstraction (Lord, 1980). Rather, parallelism should be viewed as a continuum rather than a discrete end. It is possible that psychometric methods of equating (e.g., equipercentile) may need to be applied to adjust for minor discrepancies in difficulty after forms have been administered to additional randomly equivalent groups of applicants.

The goal of the Air Force is to transition the static test forms developed here to an operational item pool suitable for computer-adaptive testing (CAT). The existing item pool is relatively small in comparison to that of a CAT-ASVAB test and generally more subject to content obsolescence. Moreover, the TIFs seen in figures 5 and 6 suggest the existing item pool is not ideal for CAT because item selection algorithms choose items based (in part) on item information, which is highly concentrated at the high end of the ability distribution. Future development efforts should focus on establishing a larger, contemporary item pool containing items that provide information along the entire ability continuum. This kind of item pool is necessary to support CAT administration and to maintain proper item exposure controls for a test that is likely to be used increasingly for selection and classification.

References

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Campbell & D.J. Knapp (Eds.) (2001), *Exploring the limits in personnel selection and classification*. NY: Lawrence Erlbaum Associates.

Defense Manpower Data Center (2014). *ASVAB Fairness Information*. Retrieved June 1, 2014, from Official site of the ASVAB website: http://officialasvab.com/fairness_res.htm

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. Applied Psychological Measurement, 35, 398-409.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). Technical review of the Armed Services Vocational Aptitude Battery (ASVAB) (FR-06-25). Alexandria, VA: Human Resources Research Organization.

Kang, C. (2011, February 17). Survey of online access finds digital divide. The Washington Post. Retrieved from [http://www.washingtonpost.com/wp-dyn/content/article/2011/02/17/AR2011021707234.html?sid=ST2011021807505] December 18, 2012.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Putka, D. J., Le, H., McCloy, R. A., Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. Journal of Applied Psychology, 93(5), 959-981.

Ree, M.J., Earles, J.A. (1992). Subtest and composite validity of ASVAB forms 11, 12, and 13 for technical training courses (AFHRL-TR-81-55). Brooks, AFB, TX: U.S. Air Force Human Resources Laboratory.

Russell, T. L., & Sellman, W. S. (2008). Review of information and communications technology literacy tests. In T.L. Russell (Chair) Measuring information and communications technology literacy. Symposium conducted at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Russell, T. L., & Sellman, W. S. (Eds.) (2009). Development and pilot testing of an information and communications technology literacy test for military enlistees: Volume 1 Final Report (FR 08-128). Alexandria, VA: Human Resources Research Organization.

Russell, T. L., & Sellman, W. S. (Eds.) (2010). Information and Communication Technology Literacy test training school validation: Phase II Final Report (FR 09-89). Alexandria, VA: Human Resources Research Organization.

Stocking, M. & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Thissen, D. (2003). MULTILOG user's guide, version 7.03 [computer program]. Chicago: Scientific Software International.

Trippe, D. M., & Russell, T. L. (Eds.) (2011). Information and communications technology literacy test norming study: Phase III final report (AFCAPS-FR-2011-00xx). Randolph AFB, TX. Air Force Personnel Center.

van der Linden, W. J. (2005). Linear models for optimal test assembly. New York, NY: Springer.

Waters, S. D. (2008). Guidelines for item sensitivity and bias review. Alexandria, VA: Human Resources Research Organization.

Welsh, J.R., Jr., Kucinkas, S.K., & Curran, L.T. (1990). Armed Services Vocational Aptitude Battery (ASVAB): Integrative review of validity studies (AFHRL-TR-90-22). Brooks AFB, TX: U.S. Air Force Human Resources Laboratory.